Connecting context, statistics and software for understanding a randomization test: a case study

Susanne Podworny

Connecting context, statistics and software for understanding a randomization test: a case study

Susanne Podworny

Paderborn University, Paderborn, Germany; podworny@math.upb.de

Drawing statistical inference with the help of simulations has gained prominence in statistics education as well as introducing statistical inference with randomization tests. This paper describes some selected results of a case study of preservice primary teachers who attended a short learning trajectory on statistical inference with randomization tests. It will be shown how the participants address the context, statistics and software level when conducting a randomization test with software and how the conscious linking of the three levels can support the learning process and help to understand certain elements of a randomization test.

Keywords: Statistical inference, randomization test, simulation.

Introduction

Statistical reasoning is a cornerstone on which statistical practice is based. In almost all areas of daily life, data and thus also conclusions drawn from data play an important role. It is impossible to imagine statistical practice without computer-supported evaluations and methods. In many fields, like in industry, medicine, or politics, decisions are increasingly being made on the basis of data. When looking at a newspaper, a television report or an entry on the World Wide Web, interested citizens increasingly come across the keywords "A study has shown ..." or "The effect of X is Y". However, it is often suggested that the results and interpretations delivered in this way are by no means certain, as is often suggested in the media.

There are two big areas in inferential statistics: parameter estimation and hypothesis testing and there are two kinds of inferences that may be drawn: generalizations beyond a given sample and causation for a given treatment for a given sample. The latter should be an integral part of stochastic education, as called for in the basic article by Wild and Pfannkuch (1999).

> Statistics education should really be telling students something every scientist knows, 'The quest for causes is the most important game in town.' It should be saying: 'Here is how statistics helps you in that quest'. (Wild & Pfannkuch, 1999, p. 238)

One statistical method for drawing causal inferences is the randomization method. About ten years ago, Cobb (2007) strongly advocated introducing the logic of inference via randomization tests. A randomization test is a non-parametric method that allows easy access to inferential reasoning via computer-based simulations. Through simulations, nearly no formulas or calculations are needed, and this is one of the main reasons for the easiness of this method. The "core logic of inference" (Cobb, 2007) can be in the center and conclusions are possible even for data from small or non-random samples.

From this perspective, a learning trajectory on inferential reasoning with randomization tests was developed by the author to be implemented in an existing course on statistics and probability for preservice primary school teachers to complete the general statistics education of these preservice

teachers. Eight weeks after the learning trajectory a case study was conducted with participants who conducted a randomization test for a given problem with computer-based simulations. The objective of this paper is to better understand the reasoning process of these learners. Selected findings of this study will be presented in this paper.

Literature review

Cobb (2007) gave an impetus to rethink the introduction to inference statistics, especially at college level, and to get a new introduction with randomization tests. Some curricula for introductory statistics courses emerged since then (e.g. Rossman, Chance, Cobb, & Holcomb, 2008; Tintle, VanderStoep, & Swanson, 2009; Zieffler & Catalysts for Change, 2013) and some shorter learning trajectories for introducing inferential reasoning were created (e.g. (Budgett, Pfannkuch, Regan, & Wild, 2012; Frischemeier & Biehler, 2014). All these teaching proposals are based on the use of computer-based simulations and focus on the logic of inferential reasoning rather than on calculations. In addition, there are some few empirical studies focusing on special aspects of the process when learners conduct a randomization test.

A common factor of all these learning units is that they use a plan or a scheme to structure the reasoning process. A compilation of elements of these schemes by the author has resulted in nine elements that can be considered central when conducting a randomization test. The first element is the random allocation of experimental units to groups. Explaining this is a core component in understanding the underlying design (Pfannkuch, Budgett, & Arnold, 2015). To find possible explanations for observed differences between two groups of an experiment (Pfannkuch et al., 2015) is the second element. One explanation can be that the treatment is effective, another explanation can be that the observed differences are due to the random allocation of units to the groups. The third element is to pose or reconstruct the research question for the experiment (Wild & Pfannkuch, 1999). Analyzing the observed data and identifying a difference between the two groups is the fourth element (Biehler, Frischemeier, & Podworny, 2015; Rossman et al., 2008). The fifth element is setting up the null model with the null and alternative hypotheses (Biehler et al., 2015). Transferring the null model to a simulation model is the sixth element (Biehler et al., 2015; Lee, Tran, Nickell, & Doerr, 2015). The seventh element is the production of test statistics and the sampling distribution (Lee et al., 2015). The eighth element is to identify the *p*-value (Biehler et al., 2015; Rossman et al., 2008). And the last element is drawing possible conclusions (Cobb, 2007). Each of these elements has its own difficulties, but one difficulty across all elements is to combine the levels of context, statistics and software.

In our own research (Biehler et al., 2015), we identified three levels when working on a randomization test: the context level, the statistics level, and the software level. Following up on this, Noll and Kirin (2017) conducted a case study with eight learners who created different models in TinkerPlots for the "dolphin therapy problem" (Noll & Kirin, 2017, p. 219) and analyzed what influenced the learners' reasoning about models. One of their results was, that "Students did not spend much time discussing attribute labels or what type of devices they wanted to use" (Noll & Kirin, 2017, p. 232). This may be an explanation for the reported result that "the concept of no difference between two groups is difficult to operationalize into a TinkerPlots model" (Noll & Kirin, 2017, p. 235). Naming

the (simulated) attributes was identified by Noll and Kirin as an important aspect of the modeling process, because it can act as a bridge between the context and the tool (Noll & Kirin, 2017, p. 236). But they also state, that they "see this work as in its infancy in that we need more research focused on why students create the models" (Noll & Kirin, 2017, p. 240). They also noted that it is not reasonable to separate the three worlds like they interpreted the work of (Biehler et al., 2015). At this connection between context, statistics and software this paper ties in.

The learning trajectory "Inferential reasoning with randomization tests"

A learning trajectory "Inferential reasoning with randomization tests" for preservice primary school teachers was created in the sense of the design based research paradigm (Cobb, Confrey, diSessa, Lehrer, & Schauble, 2003). The aim of the trajectory was to introduce preservice teachers at university to the logic of inference like in the sense of Cobb (2007) in a short period of time. Some design ideas were adapted from Pfannkuch et al. (2015), who successfully introduced inferential reasoning in a short learning sequence to a similar target group. The learning trajectory was designed for the end of an existing course on elementary statistics at Paderborn University. The existing course consisted of three modules data analysis, combinatorics, and probability and used the software TinkerPlots from the beginning. The new learning trajectory was designed for three 90 minutes sessions.

In the first session the students should get in touch with a first example of a randomization test with a significant *p*-value (Rossman et al., 2008) and with the logic of inferential reasoning as a continuation of group comparisons (Makar & Confrey, 2002), which were already discussed in the data analysis module. The second session focused on the performance of a randomization test by the students. First, a hands on simulation with pen and paper was to be carried out and then transferred to a computer-based simulation, like proposed for example by Gould, Davis, Patel, and Esfandiari (2010). In the third session, the nine elements (see literature review) were to be discussed in detail and possible difficulties addressed.

To get insights into the cognitive processes of students conducting a randomization test with TinkerPlots, an interview study was designed as part of the Ph.D. study of the author.

Methodology

The research question for the case study – that is focused on in this paper – is *How do the participants relate the three levels context, statistics and software to each other?*

The problem given to students was an adaption of the "Fish oil and blood pressure task" of Pfannkuch et al. (2015). This task contains real data from a medical experiment with 14 volunteers on the question of whether fish oil supplements have a blood pressure-lowering effect compared to normal oil supplements. The blood pressure of the participants was measured at the beginning and after four weeks and the blood pressure reductions for the two groups were recorded like in Table 1.

Fish oil group	8	12	10	14	2	0	0
"normal oil" group	-6	0	1	2	-3	-4	2

Table 1: Data on blood pressure reduction after four weeks

The observed data was visualized on the worksheet for the students like in Figure 1 and accompanied by the statement and the question "The observed data are shown in Figure 1 and show that the blood pressure reduction in the fish oil group tends to be greater than those in the 'normal oil' group". What can be concluded here?"

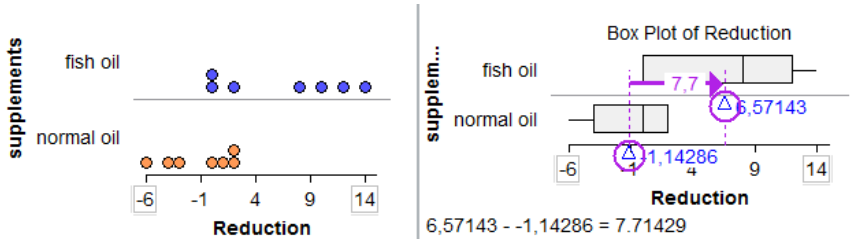

Figure 1: Visualized data on the participants' worksheet

The interview study took place as a semi structured interview (Mayring, 2016) with a large part of the participants working independently on the task. Questions of the participants were allowed in this phase, and some help could be given by the interviewer. This part was followed by interview questions relating to the individual steps and arguments of the working phase.

A total of six participants working in three pairs took part in the interview study. All of them attended the whole course and the learning trajectory, so this was their educational background. Participation in the study was voluntary, so participants cannot be considered representative of the 236 participants in the whole course. The participants' age ranged between 21 and 25 years, being in the third/fourth semester of University. The interviews took place eight weeks after the course and none of the participants used TinkerPlots in between. The conversations were recorded together with the screen activities and then transferred into a transcript. This transcript included the conversations as well as the other activities, with and without software.

The analysis of the transcripts was carried out by means of interaction analysis (Krummheuer & Naujok, 1999). First, the transcripts were divided into 15 interaction units using methods of linguistic conversation analysis (Egbert & Deppermann, 2012). The next step was to reconstruct the solution process interpretatively with interaction analysis. A detailed turn-by-turn analysis took place here and was discussed with other interpreters. For each of the nine elements (see literature review), the level (context, statistics, software) at which the participants communicated was examined. As a third step, the use of the software in the solution process was examined in detail. Summary and comparative analyses between the pairs formed the fourth step.

Results

The levels at which the participants communicate regarding the nine elements are described below. Table 2 shows which level the participants address linguistically when they talk about and work on the various elements.

Element	Context	Statistics	Software
Random allocation	Rebecca & Selina Fabia & Laura Mandy & Alisa	Rebecca & Selina Fabia & Laura Mandy & Alisa	- - -
Possible explanation	Rebecca & Selina Fabia & Laura	Rebecca & Selina Fabia & Laura	- -

	Mandy & Alisa	Mandy & Alisa	-
Research question	Rebecca & Selina Fabia & Laura Mandy & Alisa	- - -	- - -
Observed data	Rebecca & Selina Fabia & Laura Mandy & Alisa	Rebecca & Selina Fabia & Laura Mandy & Alisa	- - -
Null model	Rebecca & Selina Fabia & Laura Mandy & Alisa	Rebecca & Selina Fabia & Laura Mandy & Alisa	- - -
Simulation model	Rebecca & Selina Fabia & Laura Mandy & Alisa	Rebecca & Selina Fabia & Laura Mandy & Alisa	Rebecca & Selina Fabia & Laura Mandy & Alisa
Test statistics and sampling distribution	- - -	Rebecca & Selina Fabia & Laura Mandy & Alisa	Rebecca & Selina Fabia & Laura Mandy & Alisa
P-value	- - -	Rebecca & Selina Fabia & Laura Mandy & Alisa	Rebecca & Selina Fabia & Laura Mandy & Alisa
Conclusions	- - Mandy & Alisa	Rebecca & Selina Fabia & Laura Mandy & Alisa	- - -

Table 2: Overview of the levels at which the pairs communicate during the randomization test process

Table 2 shows a clear pattern. The first two elements are addressed by all three pairs both at the context level and at the statistics level. These are related to the observed data and no reference to the software is necessary. The research question is formulated by the participants at the contextual level. The corresponding interaction units are very short, because this element is obviously clear for all participants. Like the first two elements, the observed data are discussed at the contextual and at the statistical level. As the data had already been evaluated on the worksheet, there was no need to work with the software and therefore no need to talk about it.

The null model element includes the formulation of the null hypothesis and the alternative hypothesis. Communication takes place both at the context level and at the statistics level. However, the null hypothesis formulated by Rebecca and Selina reveals only a small contextual reference. These two formulate “Random group allocation is the cause of observed differences”. Such a formulation is almost arbitrarily applicable to different situations, since a reference to the direct context is not recognizable. The formulation of the other pairs for the null hypothesis are “There is no difference in blood pressure reduction in the effect between both supplements” (Fabia and Laura) and “It does not matter which oil is taken to lower the blood pressure but the results are due to the random allocation” (Mandy and Alisa). Both show a clear connection to the context. Perhaps this explains why Rebecca and Selina do not end up drawing conclusions in context. However, Fabia and Laura do not draw these either, although their null hypothesis is clearly related to the context.

The only element that is addressed by all three pairs on all three levels is the development of the simulation model in TinkerPlots. For this element, all three pairs needed additional help from the interviewer on a technical level, for example how to copy all the observed values into a box of the sampler. The same need of help could be observed for the next two elements for which TinkerPlots

had to be used. But, unlike in our earlier research (Biehler et al., 2015), renaming the attributes of the sampler and of the new groups for the new allocation was a topic for all three pairs. An interesting dialogue for this came about between Rebecca and Selina. This shows that the concept of random allocation developed linguistically by linking the three levels (that was identified as a difficulty by Noll & Kirin (2017)). The sampler built by Rebecca (R) and Selina (S) so far is shown in Figure 2. This sampler is built correctly with all observed blood pressure values in the first box and independently these will be drawn in one of two new groups. TinkerPlots labelled the two stacks in the second device in "a" and "b" automatically.

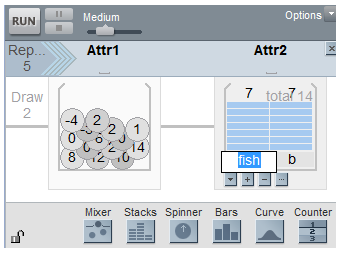

Figure 2: Sampler created by Rebecca and Selina

Selina first renames the first stack in "fish" like in Figure 2, but then retains the letters a and b and the following dialogue arises.

R: Do we have to call them that, because we actually divide them up in such a way that we can do it later, for example/. So I'm not quite sure, but it could be that we/.

S: /That's right, we don't really need it.
R: /put a fish oil person with a normal oil person/.
S: /You're right/.
R: /Together. I think a and b is actually quite neutral.
S: You can actually take a and b, because then that's more. We want to prove now that it has nothing to do with the oil anymore.
R: Yes, exactly we will say it/.
S: /So you have to/
R: We are representatives of the null hypothesis.
S: Exactly, because then in the one group possibly always both come, a person with fish oil and one with normal oil. It is now only about lowering blood pressure.
R: Yes, that's exactly what it can be/.
S: /Yes.
R: That you have maybe five of them *(points to the fish oil group in the plot in TinkerPlots)* and then (.) the rest of them *(points to the "normal oil" group in the plot)*. Well, it can be that way/.
S: /You're right. *(laughs)* Well, I would spontaneously agree with you. *(laughs)*

In the end, Rebecca and Selina decide not to rename the new groups.

For the next two elements, sampling distribution and *p*-value, none of the three pairs have a conversation at context level. They operate only on the software level and on the statistics level.

Conclusions are correctly formulated at the statistical level by all three pairs, which is a good result for the whole process. The written formulations are as follows:

Rebecca & Selina: Research hypothesis can be accepted and null hypothesis can be rejected, but with slight uncertainty

Fabia & Laura: We reject the null hypothesis and accept the research hypothesis at a P-value of 1%.

Mandy & Alisa: The null hypothesis can be rejected. Blood pressure reduction depends on the type of oil.

However, only Mandy and Alisa apply their conclusion to the treatment carried out. A deeper look at the transcripts shows that the greatest difficulties are with the last element formulating contextual conclusions, and that a statistical formulation seems to be the easiest for the participants.

Discussion

Like Noll and Kirin (2017), who did not find a separation of the three levels context, statistics, and software to be beneficial, connecting the three levels seems to be extremely useful in the present work. Only when the levels are differentiated a conscious connection between them can take place and thus promote the learning process. An example for a helpful connection can be seen in the selected dialogue between Rebecca and Selina, in which they clarify the meaning of random assignment and its implementation in the software.

For the process of understanding it can be concluded that the dialogue between Rebecca and Selina is one of the most important parts of the conversation, because here the two make it clear that during the simulation (level of software) a random reallocation (level of statistics) of the persons with the blood pressure values (level of context) happens independently (level of statistics) of the oil consumed (level of context). Such discussion would have been desirable for further settings at the software level, as it would enhance understanding by linking the three levels. Unfortunately, very few such dialogues have taken place throughout the case studies. Even though the data are taken from a real scientific experiment, the participants seem to have been given too little opportunity to deal with the context in detail. This would be a better prerequisite for a deeper interpretation of the context.

However, the context caused a confrontation with the data, as reported by Gonzáles (2015), in order to promote understanding of the need of decision-making in situations with uncertainty. The demands of Cobb (2007) were implemented here in a learning trajectory and the learners were successfully introduced to statistical inference.

References

Biehler, R., Frischemeier, D., & Podworny, S. (2015). Preservice teachers' reasoning about uncertainty in the context of randomization tests. In A. Zieffler & E. Fry (Eds.), *Reasoning about uncertainty: Learning and teaching informal inferential reasoning*. Minneapolis, Minnesota: Catalyst Press.

Budgett, S., Pfannkuch, M., Regan, M., & Wild, C. (2012). *Dynamic visualizations for inference*. Paper presented at the The International Association for Statistical Education Roundtable Conference: Technology in statistics education: Virtualities and Realities, Cebu City, The Philippines.

Cobb, G. (2007). The Introductory Statistics Course: A Ptolemaic Curriculum? *Technology Innovations in Statistics Education, 1*(1), 1–15. doi:https://escholarship.org/uc/item/6hb3k0nz

Cobb, P., Confrey, J., diSessa, A., Lehrer, R., & Schauble, L. (2003). Design Experiments in Educational Research. *Educational Researcher, 32*(1), 9–13.

Egbert, M., & Deppermann, A. (2012). Introduction to conversation analysis with examples from audiology. In M. Egbert & A. Deppermann (Eds.), *Hearing aids communication. Integrating social interaction, audiology and user centered design to improve communication with hearing loss and hearing technologies* (pp. 40–47). Mannheim: Verlag für Gesprächsforschung.

Frischemeier, D., & Biehler, R. (2014). Design and exploratory evaluation of a learning trajectory leading to do randomization tests facilitated by TinkerPlots. *Proceedings of the Eighth Congress of the European Society for Research in Mathematics Education*, 799–809.

Gonzáles, O. (2015). Mathematics teachers' conceptions of how to promote decision-making while teaching statistics: The case of Japanese secondary school teachers. In K. Krainer & N. Vondrová (Eds.), *Proceedings of the Ninth Congress of the European Society for Research in Mathematics Education (Cerme9, 4-8 February 2015)* (pp. 665–671). Prague, Czech Republic: Charles University in Prague, Faculty of Education and ERME.

Gould, R., Davis, G., Patel, R., & Esfandiari, M. (2010). *Enhancing conceptual understanding with data driven labs*. Paper presented at the The Eighth International Conference on Teaching Statistics, Ljubljana, Slovenia.

Krummheuer, G., & Naujok, N. (1999). *Grundlagen und Beispiele Interpretativer Unterrichtsforschung*. Opladen: Leske+Budrich.

Lee, H., Tran, D., Nickell, J., & Doerr, H. (2015). Simulation approaches for informal inference: Models to develop understanding. In K. Krainer & N. Vondrová (Eds.), *Proceedings of the Ninth Congress of European Society for Research in Mathematics Education (Cerme9, 4-8 February 2015)* (pp. 707–713). Prague, Czech Republic: Charles University in Prague, Faculty of Education and ERME.

Makar, K., & Confrey, J. (2002). *Comparing Two Distributions: Investigating Secondary Teachers' Statistical Thinking*. Paper presented at the Sixth International Conference on Teaching Statistics, Cape Town, South Africa.

Mayring, P. (2016). *Einführung in die qualitative Sozialforschung: eine Anleitung zu qualitativem Denken* (6., überarbeitete ed.). Weinheim: Beltz.

Noll, J., & Kirin, D. (2017). TinkerPlots model construction approaches for comparing two groups: Student perspectives. *Statistics Education Research Journal, 16*(2), 213–243.

Pfannkuch, M., Budgett, S., & Arnold, P. (2015). Experiment-to-causation inference: Understanding causality in a probabilistic setting. In A. Zieffler & E. Fry (Eds.), *Reasoning about uncertainty: Learning and teaching informal inferential reasoning* (pp. 95–128). Minneapolis, Minnesota: Catalyst Press.

Rossman, A., Chance, B., Cobb, G., & Holcomb, R. (2008). *Concepts of statistical inference: Approach, scope, sequence and format for an elementary permutation-based first course*. Retrieved from http://statweb.calpoly.edu/bchance/csi/CSIcurriculumMay08.doc

Tintle, N., VanderStoep, J., & Swanson, T. (2009). *An active approach to statistical inference, preliminary edition*. Holland, Michigan: Hope College Publishing.

Wild, C., & Pfannkuch, M. (1999). Statistical Thinking in Empirical Enquiry. *International Statistical Review, 67*(3), 223-265.

Zieffler, A., & Catalysts for Change. (2013). *Statistical Thinking. A simulation approach to modeling uncertainty* (3. edition ed.). Minneapolis, Minnesota: Catalyst Press.